LASSO Regression: Sparsity through L1 Regularization

Nipun Batra

IIT Gandhinagar

July 30, 2025

Outline

Ridge vs LASSO: The Key Difference

Ridge vs LASSO: The Key Difference

Ridge vs LASSO: The Key Difference

Ridge Regression

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_2^2$

L2 penalty: $\|\boldsymbol{\theta}\|_2^2 = \sum_j \theta_j^2$
Effect: Shrinks coefficients toward zero but never exactly zero

Ridge vs LASSO: The Key Difference

Ridge Regression

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_2^2$

L2 penalty: $\|\boldsymbol{\theta}\|_2^2 = \sum_j \theta_j^2$
Effect: Shrinks coefficients toward zero but never exactly zero

Ridge vs LASSO: The Key Difference

Ridge Regression

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_2^2$

L2 penalty: $\|\boldsymbol{\theta}\|_2^2 = \sum_j \theta_j^2$
Effect: Shrinks coefficients toward zero but never exactly zero

Ridge vs LASSO: The Key Difference

Ridge Regression

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_2^2$

L2 penalty: $\|\theta\|_2^2 = \sum_j \theta_j^2$
Effect: Shrinks coefficients toward zero but never exactly zero

LASSO Regression

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$

L1 penalty: $\|\theta\|_1 = \sum_j |\theta_j|$
Effect: Can set coefficients exactly to zero (sparsity!)

Ridge vs LASSO: The Key Difference

Ridge Regression

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_2^2$

L2 penalty: $\|\theta\|_2^2 = \sum_j \theta_j^2$
Effect: Shrinks coefficients toward zero but never exactly zero

LASSO Regression

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$

L1 penalty: $\|\theta\|_1 = \sum_j |\theta_j|$
Effect: Can set coefficients exactly to zero (sparsity!)

Ridge vs LASSO: The Key Difference

Ridge Regression

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_2^2$

L2 penalty: $\|\boldsymbol{\theta}\|_2^2 = \sum_j \theta_j^2$
Effect: Shrinks coefficients toward zero but never exactly zero

LASSO Regression

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$

L1 penalty: $\|\boldsymbol{\theta}\|_1 = \sum_j |\theta_j|$
Effect: Can set coefficients exactly to zero (sparsity!)

The Magic of L1

LASSO performs **automatic feature selection** by setting irrelevant coefficients to zero!

Pop Quiz: Ridge vs LASSO

Quick Quiz 1

Which method is better for feature selection?

a) Ridge Regression (L2 penalty)

Answer: b) LASSO can set coefficients exactly to zero, effectively removing features from the model!

Pop Quiz: Ridge vs LASSO

Quick Quiz 1

Which method is better for feature selection?

a) Ridge Regression (L2 penalty)
b) LASSO Regression (L1 penalty)

Answer: b) LASSO can set coefficients exactly to zero, effectively removing features from the model!

Pop Quiz: Ridge vs LASSO

Quick Quiz 1

Which method is better for feature selection?

a) Ridge Regression (L2 penalty)
b) LASSO Regression (L1 penalty)
c) Both are equally good for feature selection

Answer: b) LASSO can set coefficients exactly to zero, effectively removing features from the model!

LASSO Optimization Problem

LASSO Objective

$$\text{Minimize } \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{d} |\theta_j|$$

LASSO Optimization Problem

LASSO Objective

$$\text{Minimize } \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{d} |\theta_j|$$

LASSO Optimization Problem

LASSO Objective

$$\text{Minimize } \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{d} |\theta_j|$$

Challenge: The L1 penalty $|\theta_j|$ is not differentiable at $\theta_j = 0$!

LASSO Optimization Problem

LASSO Objective

$$\text{Minimize } \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{d} |\theta_j|$$

Challenge: The L1 penalty $|\theta_j|$ is not differentiable at $\theta_j = 0$!

Solution approaches:

- Coordinate Descent: Update one coefficient at a time

LASSO Optimization Problem

LASSO Objective

$$\text{Minimize } \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{d} |\theta_j|$$

Challenge: The L1 penalty $|\theta_j|$ is not differentiable at $\theta_j = 0$!

Solution approaches:

- Coordinate Descent: Update one coefficient at a time
- Subgradient Methods: Use subderivatives

LASSO Optimization Problem

LASSO Objective

$$\text{Minimize } \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{d} |\theta_j|$$

Challenge: The L1 penalty $|\theta_j|$ is not differentiable at $\theta_j = 0$!

Solution approaches:

- Coordinate Descent: Update one coefficient at a time
- Subgradient Methods: Use subderivatives
- Proximal Methods: Soft thresholding operators

Coordinate Descent Strategy

Key Idea: Update one coefficient θ_j while keeping others fixed

Partial Prediction

$$\hat{y}_i^{(-j)} = \sum_{k \neq j} \theta_k x_{ik} = \hat{y}_i - \theta_j x_{ij}$$

Interpretation: Prediction without the j-th feature's contribution

Coordinate Descent Strategy

Key Idea: Update one coefficient θ_j while keeping others fixed

Partial Prediction

$$\hat{y}_i^{(-j)} = \sum_{k \neq j} \theta_k x_{ik} = \hat{y}_i - \theta_j x_{ij}$$

Interpretation: Prediction without the j-th feature's contribution

Coordinate Descent Strategy

Key Idea: Update one coefficient θ_j while keeping others fixed

Partial Prediction

$$\hat{y}_i^{(-j)} = \sum_{k \neq j} \theta_k x_{ik} = \hat{y}_i - \theta_j x_{ij}$$

Interpretation: Prediction without the j-th feature's contribution

Residual for coordinate j:

$$r_j = y_i - \hat{y}_i^{(-j)} = y_i - \sum_{k \neq j} \theta_k x_{ik}$$

Coordinate Descent Strategy

Key Idea: Update one coefficient θ_j while keeping others fixed

Partial Prediction

$$\hat{y}_i^{(-j)} = \sum_{k \neq j} \theta_k x_{ik} = \hat{y}_i - \theta_j x_{ij}$$

Interpretation: Prediction without the j-th feature's contribution

Residual for coordinate j:

$$r_j = y_i - \hat{y}_i^{(-j)} = y_i - \sum_{k \neq j} \theta_k x_{ik}$$

The Update Problem

Find θ_j that minimizes: $\sum_i (r_j - \theta_j x_{ij})^2 + \lambda |\theta_j|$

Soft Thresholding: The LASSO Solution

For unregularized regression: $\theta_j = \frac{\rho_j}{z_j}$

where:

$$\rho_j = \sum_{i=1}^{n} x_{ij}(y_i - \hat{y}_i^{(-j)}) \text{ (correlation with residual)} \tag{1}$$

$$z_j = \sum_{i=1}^{n} x_{ij}^2 \text{ (squared norm of feature j)} \tag{2}$$

Soft Thresholding: The LASSO Solution

For unregularized regression: $\theta_j = \frac{\rho_j}{z_j}$

where:

$$\rho_j = \sum_{i=1}^{n} x_{ij}(y_i - \hat{y}_i^{(-j)}) \text{ (correlation with residual)} \tag{1}$$

$$z_j = \sum_{i=1}^{n} x_{ij}^2 \text{ (squared norm of feature j)} \tag{2}$$

LASSO Update (Soft Thresholding)

$$\theta_j = \begin{cases} \frac{\rho_j - \lambda}{z_j} & \text{if } \rho_j > \lambda \\ 0 & \text{if } |\rho_j| \leq \lambda \\ \frac{\rho_j + \lambda}{z_j} & \text{if } \rho_j < -\lambda \end{cases}$$

Soft Thresholding: The LASSO Solution

For unregularized regression: $\theta_j = \frac{\rho_j}{z_j}$

where:

$$\rho_j = \sum_{i=1}^{n} x_{ij}(y_i - \hat{y}_i^{(-j)}) \text{ (correlation with residual)} \tag{1}$$

$$z_j = \sum_{i=1}^{n} x_{ij}^2 \text{ (squared norm of feature j)} \tag{2}$$

LASSO Update (Soft Thresholding)

$$\theta_j = \begin{cases} \frac{\rho_j - \lambda}{z_j} & \text{if } \rho_j > \lambda \\ 0 & \text{if } |\rho_j| \leq \lambda \\ \frac{\rho_j + \lambda}{z_j} & \text{if } \rho_j < -\lambda \end{cases}$$

Soft Thresholding: The LASSO Solution

For unregularized regression: $\theta_j = \frac{\rho_j}{z_j}$

where:

$$\rho_j = \sum_{i=1}^{n} x_{ij}(y_i - \hat{y}_i^{(-j)}) \text{ (correlation with residual)} \tag{1}$$

$$z_j = \sum_{i=1}^{n} x_{ij}^2 \text{ (squared norm of feature j)} \tag{2}$$

LASSO Update (Soft Thresholding)

$$\theta_j = \begin{cases} \frac{\rho_j - \lambda}{z_j} & \text{if } \rho_j > \lambda \\ 0 & \text{if } |\rho_j| \leq \lambda \\ \frac{\rho_j + \lambda}{z_j} & \text{if } \rho_j < -\lambda \end{cases}$$

Key insight: If correlation $|\rho_j|$ is small, set $\theta_j = 0$

Pop Quiz: Soft Thresholding

Quick Quiz 2

In LASSO soft thresholding, what happens when $|\rho_j| \leq \lambda$?

a) θ_j becomes very large

Answer: b) When correlation with residual is small, LASSO sets the coefficient to zero!

Pop Quiz: Soft Thresholding

Quick Quiz 2

In LASSO soft thresholding, what happens when $|\rho_j| \leq \lambda$?

a) θ_j becomes very large
b) θ_j is set exactly to zero

Answer: b) When correlation with residual is small, LASSO sets the coefficient to zero!

Pop Quiz: Soft Thresholding

Quick Quiz 2

In LASSO soft thresholding, what happens when $|\rho_j| \leq \lambda$?

a) θ_j becomes very large
b) θ_j is set exactly to zero
c) θ_j remains unchanged

Answer: b) When correlation with residual is small, LASSO sets the coefficient to zero!

Ridge vs LASSO: When to Use Which?

Ridge vs LASSO: When to Use Which?

Ridge vs LASSO: When to Use Which?

Use Ridge When:

Result: All features kept with small coefficients

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features

Result: All features kept with small coefficients

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients

Result: All features kept with small coefficients

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues

Result: All features kept with small coefficients

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues
- Smooth solutions preferred

Result: All features kept with small coefficients

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues
- Smooth solutions preferred

Result: All features kept with small coefficients

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues
- Smooth solutions preferred

Result: All features kept with small coefficients

Use LASSO When:

Result: Only important features kept

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues
- Smooth solutions preferred

Result: All features kept with small coefficients

Use LASSO When:

- Many irrelevant features

Result: Only important features kept

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues
- Smooth solutions preferred

Result: All features kept with small coefficients

Use LASSO When:

- Many irrelevant features
- Want automatic feature selection

Result: Only important features kept

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues
- Smooth solutions preferred

Result: All features kept with small coefficients

Use LASSO When:

- Many irrelevant features
- Want automatic feature selection
- Sparse solutions desired

Result: Only important features kept

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues
- Smooth solutions preferred

Result: All features kept with small coefficients

Use LASSO When:

- Many irrelevant features
- Want automatic feature selection
- Sparse solutions desired
- Interpretability important

Result: Only important features kept

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues
- Smooth solutions preferred

Result: All features kept with small coefficients

Use LASSO When:

- Many irrelevant features
- Want automatic feature selection
- Sparse solutions desired
- Interpretability important

Result: Only important features kept

Ridge vs LASSO: When to Use Which?

Use Ridge When:

- Many relevant features
- Want to shrink coefficients
- Multicollinearity issues
- Smooth solutions preferred

Result: All features kept with small coefficients

Use LASSO When:

- Many irrelevant features
- Want automatic feature selection
- Sparse solutions desired
- Interpretability important

Result: Only important features kept

Best of Both Worlds

Elastic Net: Combines L1 and L2 penalties:
$\alpha\|\boldsymbol{\theta}\|_1 + (1-\alpha)\|\boldsymbol{\theta}\|_2^2$

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly
  - Never exactly zero coefficients

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly
  - Never exactly zero coefficients
  - Good when all features are somewhat relevant

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly
    - Never exactly zero coefficients
    - Good when all features are somewhat relevant

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly
    - Never exactly zero coefficients
    - Good when all features are somewhat relevant
- **LASSO Solution:** L1 penalty enables feature selection

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly
  - Never exactly zero coefficients
  - Good when all features are somewhat relevant
- **LASSO Solution:** L1 penalty enables feature selection
  - Can set coefficients exactly to zero

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly
  - Never exactly zero coefficients
  - Good when all features are somewhat relevant
- **LASSO Solution:** L1 penalty enables feature selection
  - Can set coefficients exactly to zero
  - Automatic feature selection

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly
  - Never exactly zero coefficients
  - Good when all features are somewhat relevant
- **LASSO Solution:** L1 penalty enables feature selection
  - Can set coefficients exactly to zero
  - Automatic feature selection
  - Solved via soft thresholding

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly
  - Never exactly zero coefficients
  - Good when all features are somewhat relevant
- **LASSO Solution:** L1 penalty enables feature selection
  - Can set coefficients exactly to zero
  - Automatic feature selection
  - Solved via soft thresholding

Summary: Regularization Methods

- **Problem:** Overfitting in linear regression
- **Ridge Solution:** L2 penalty shrinks coefficients smoothly
  - Never exactly zero coefficients
  - Good when all features are somewhat relevant
- **LASSO Solution:** L1 penalty enables feature selection
  - Can set coefficients exactly to zero
  - Automatic feature selection
  - Solved via soft thresholding
- **Key Insight:** Choice depends on problem structure and interpretability needs

(2)

$$\frac{\partial}{\partial \theta_j} |\theta_j| = \begin{cases} 1 & \theta_j > 0 \\ [-1, 1] & \theta_j = 0 \\ -1 & \theta_j < 0 \end{cases}$$

Coordinate Descent for Lasso Regression

- **Case 1:** $\theta_j > 0$

$$-2\rho_j + 2\theta_j z_j + \delta^2 = 0$$

$$\theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

$$\rho_j > \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

Coordinate Descent for Lasso Regression

- **Case 1:** $\theta_j > 0$

$$-2\rho_j + 2\theta_j z_j + \delta^2 = 0$$

$$\theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

$$\rho_j > \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

Coordinate Descent for Lasso Regression

- **Case 1:** $\theta_j > 0$

$$-2\rho_j + 2\theta_j z_j + \delta^2 = 0$$

$$\theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

$$\rho_j > \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

- **Case 2:** $\theta_j < 0$

$$\rho_j < \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j + \delta^2/2}{z_j} \quad (3)$$

Coordinate Descent for Lasso Regression

- **Case 3:** $\theta_j = 0$

$$\frac{\partial}{\partial \theta_j}(\text{LASSO OBJECTIVE}) = -2\rho_j + 2\theta_j z_j + \delta^2 \underbrace{\frac{\partial}{\partial \theta_j} |\theta_j|}_{\text{[-1,1]}}$$

$$\in \underbrace{[-2\rho_j - \delta^2, -2\rho_j + \delta^2]}_{\{0\} \text{ lies in this range}}$$

$$-2\rho_j - \delta^2 \leq 0 \text{ and } -2\rho_j + \delta^2 \geq 0$$

$$-\frac{\delta^2}{2} \leq \rho_j \leq \frac{\delta^2}{2} \Rightarrow \theta_j = 0$$

Summary of Lasso Regression

$$\theta_j = \left[\begin{array}{ccc} \frac{\rho_j + \frac{\delta^2}{2}}{z_j} & \text{if} & \rho_j < -\frac{\delta^2}{2} \\ 0 & \text{if} & -\frac{\delta^2}{2} \leq \rho_j \leq \frac{\delta^2}{2} \\ \frac{\rho_j - \frac{\delta^2}{2}}{z_j} & \text{if} & \rho_j > \frac{\delta^2}{2} \end{array} \right] \tag{4}$$